Journal of the American College of Cardiology
© 2001 by the American College of Cardiology
Published by Elsevier Science Inc.

Vol. 37, No. 1, 2001
ISSN 0735-1097/01/$20.00
PII S0735-1097(00)01082-2

EDITORIAL COMMENT

Testing Our Tests: Surrogate End Points Versus Driving Patient Management and Outcomes*

James E. Udelson, MD, FACC
Boston, Massachusetts

Ideally, in choosing cardiovascular therapeutics for a specific clinical syndrome, we make evidence-based, outcome-driven decisions informed by the results of randomized prospective clinical trials when such information is available. This type of information forms the most robust basis of disease-based practice guidelines, which also summarize less robust types of evidence, such as observational data from large trials as well as expert opinion.

See page 81

When making decisions about choosing a particular testing modality to assess a disease process in order to make a clinical management decision, we far less often have such decisions informed by evidence from randomized clinical trials. In the case of various types of stress testing, we have observational data assessing test performance in relation to a "gold standard," such as a 50% or more stenosis by angiography, as well as numerous studies examining the association between stress-testing results and subsequent outcomes for prognosis. In situations where one or more distinct testing modalities may be used to assess/diagnose/prognosticate the same disease process, clinicians will often base their choice of testing modality on their own experience and local expertise, their impression of the comparative value of the modalities based on individual studies in the literature, and now perhaps on emerging meta-analyses comparing performance of different testing modalities (1,2). While the meta-analytic approach overcomes problems associated with relatively small single-center studies, how to account for such factors as verification bias and the technical evolution of testing modalities over time is challenging.

Although less common than in the therapeutic arena, prospective randomized clinical trials examining different testing modalities linked to clinical decisions are occasionally undertaken and reported. Studies such as the Thrombolysis in Myocardial Infarction (TIMI) 2 and the VANQWISH trials in acute coronary syndromes (3,4), while often referred to as comparisons of aggressive versus conservative therapeutic management strategies, are in essence randomized trials of the influence of a choice of initial testing strategy on subsequent outcome. These trials examine the strategy of initial referral directly to assessment by catheterization, with subsequent revascularization based on anatomic considerations, compared to initial assessment with stress radionuclide imaging, with subsequent catheterization/revascularization based on physiologic considerations of the extent of inducible ischemia. The results of these studies comparing initial risk-stratification testing strategies generally support that initial direct referral to catheterization will not necessarily result in better long-term outcomes.

In the clinical syndrome of chronic coronary artery disease (CAD) accompanied by left ventricular (LV) dysfunction, where issues regarding myocardial viability and the potential reversibility of left ventricular dysfunction with revascularization are important drivers of clinical decision making, several distinct testing modalities have been investigated extensively over the years. Single-photon emission computed tomography imaging, using thallium-201 or Tc-99m–based agents, examines sarcolemmal membrane integrity after tracer delivery via myocardial blood flow (5). Positron emission tomographic (PET) imaging interrogates oxidative or glucose metabolism as well as myocardial blood flow (6), while dobutamine echocardiography examines regional wall motion and the effect of inotropic stimulation on contractile reserve (7). The parameters examined by these testing modalities are characteristic features of viable but dysfunctional myocardium, a scenario in which clinical outcome may be improved by revascularization.

Evaluation of each of these testing modalities has followed a somewhat stereotypical trajectory over the years. Early reports tended to examine changes in tracer uptake itself—for example, with different iterations of thallium-201 imaging protocols, implying evidence of improved identification of dysfunctional but viable myocardium (8). The individual testing modalities were also examined for their performance characteristics in predicting recovery of regional function after revascularization (8–11). Some authors suggested that this gold standard is the sine qua non of myocardial viability assessment, as the aim should always be to improve function of dyssynergic myocardium (12). Numerous papers have reported on cross-correlative comparisons between the testing modalities, examining proportions of agreement or disagreement. Often in such studies, PET evidence of preserved metabolic activity is considered the gold standard comparator. One could question this assumption given the imperfect performance of PET imaging itself for predicting functional recovery. Changes in global ejection fraction after revascularization have been less often studied in this setting as a gold standard, but they consistently have been shown to follow revascularization of a

*Editorials published in the *Journal of the American College of Cardiology* reflect the views of the authors and do not necessarily represent the views of *JACC* or the American College of Cardiology.

From the Division of Cardiology, New England Medical Center Hospitals, Tufts University School of Medicine, Boston, Massachusetts.

JACC Vol. 37, No. 1, 2001
January 2001:89–92

certain threshold mass of viable dysfunctional myocardium (10,13).

What all of these studies have in common is the use of surrogate end points to evaluate test performance. If one takes a patient-based or outcome-based perspective, the major consideration in making a clinical decision for revascularization in the case of chronic CAD, LV dysfunction, and heart failure is whether the patient will feel better (assessed in terms of symptoms or functional capacity in some way) and/or experience an improved natural history than they might have otherwise without revascularization. How tightly these patient-based outcomes are tied to the surrogate end points often used in viability studies is not at all well established. As has been pointed out, there are many physiologic outcomes that may follow revascularization and be associated with improved clinical outcome that do not necessarily involve recovery of regional or even global systolic performance (14,15). Such features include improvement in diastolic function, improvement in arrhythmic milieu, prevention of myocardial infarction, and attenuation of remodeling. Recent data have suggested that survival after revascularization in the setting of chronic CAD and LV dysfunction in patients undergoing bypass surgery may be independent of whether ejection fraction (EF) improves or not (16). Although some data challenge this (17), much that we understand about the complex pathophysiology of chronic ischemic LV dysfunction would suggest that post-revascularization improvement in measures of systolic performance may be a sufficient but not necessary condition for improved long-term patient outcome.

This concept calls into question the use of any surrogate end point for the critical analysis for testing modalities assessing myocardial viability. For instance, cross-correlative studies will often show some advantage to one of two modalities being directly compared. Although differences in proportions of myocardial segments expressing a certain amount of tracer uptake, metabolic activity or inotropic reserve may become statistically significant, the actual total amount of myocardium per patient, or alternatively the number of patients with a significant amount of myocardium involved, is usually rather small (15,18). When the myocardium is segmented into 16 or 20 or even 40 segments, the number of data points to compare is enlarged sufficiently to do a careful segmental investigation, but statistical differences may indeed represent relatively small amounts of total LV myocardium.

Even carefully performed meta-analyses comparing performance of the different testing modalities may be of limited generalizability based on the use (by necessity) of regional functional recovery as the end point of interest (19). If one believes that regional functional recovery may be a critically important determinant of outcome, then such meta-analyses would suggest that using techniques such as SPECT imaging or dobutamine echocardiography, whose performance characteristics are not quite as robust as PET, would lead to less optimal long-term outcomes if used to drive clinical decisions in this setting. If, however, one takes the view that the amount of total myocardium represented by differences in the tests is actually quite small, then one might expect no difference in long-term outcome no matter which test is employed. Indeed, a preliminary report of a meta-analysis examining the relationship between myocardial viability testing and differential outcomes with revascularization and medical therapy reveals no differences between the testing modalities (20). Thus, the question lingers whether the relatively small differences between the testing modalities actually would result in any important differences in patient-based management or outcomes.

How best do we critically address this dilemma? A study in this issue of the *Journal* marks a very important step forward in this direction. In an elegantly designed trial meant to address these very questions, Siebelink and colleagues (21) have performed a prospective randomized clinical trial comparing management decisions and outcomes based on either PET imaging or SPECT imaging in patients with chronic CAD and LV dysfunction. By the inclusion criteria, the investigators have focused on patients in whom such testing is crucial for clinical decisions; that is, those in whom questions of regional viability will be the primary drivers of the dichotomous decision to revascularize or not. Statistically, the study was powered to detect a 20% difference in long-term outcomes between the PET-driven and SPECT-driven management, based on observational trials comparing outcomes with the two testing modalities (22). Although all patients underwent both PET and SPECT imaging, each patient was randomized to have information from only one of the modalities revealed to the clinical decision-making team, in a polar map format identifying normal, viable, and infarcted myocardium in a way that the clinicians could not recognize whether the data came from the SPECT or PET imaging. Clinical decisions were made on the basis of the data provided, and the patients were managed accordingly and followed for long-term outcomes. The results demonstrated no significant differences in the proportions of patients sent to revascularization or managed medically; most importantly, there were no differences in long-term event-free survival between the groups managed based on results from the different techniques. These data suggest that for the majority of patients with this clinical syndrome in whom revascularization is being considered, widely available SPECT imaging techniques, when performed and interpreted with expertise, can drive patient management decisions and outcomes in a manner similar to PET imaging.

There are some important features of this study to point out (21). Only about one-third of the population had very significant LV dysfunction (EF $<$ 30%), whereas in practice it is likely that a larger proportion of patients with questions of viability being considered for revascularization, particularly coronary artery bypass grafting (CABG), fall into this category. However, the investigators found no difference in outcomes even among this subgroup of patients with severe

LV dysfunction using the two different testing modalities. The PET imaging protocol used by this group of investigators is somewhat out of keeping with standard PET imaging, in that myocardial blood flow was examined during pharmacologic stress, with the stress blood flow data coupled to fluorodeoxyglucose metabolic information at rest. The information derived from SPECT Tc-99m sestamibi stress/rest imaging was categorized as normal, viable, or infarcted, and did not take full advantage of the information available on the distinction between ischemic, viable myocardium and viable myocardium without stress-induced ischemia, a difference that may be important clinically (23). The patients had a stable clinical syndrome, with the imaging and decision making taking place over 30 to 40 days; thus, these data cannot address decision making in more acute syndromes.

Finally, some might argue that the projected 20% difference in outcome, which drove the power analysis and sample size calculations, was too large, given the relatively small differences between the testing modalities in most studies. If that is true, then a type II error might have occurred, indicating that we may be concluding that there is no difference between test-driven outcomes when in fact there may be a difference. However, taking the very small and statistically nonsignificant difference found in the study and projecting forward, a sample size of close to 20,000 patients would be needed to make these differences statistically significant. Therefore, one might conclude that if a difference in long-term test-driven outcomes does exist, it is indeed quite small.

The Siebelink study makes an important contribution to the literature on myocardial viability and helps inform our choices of testing modality in a particular clinical scenario. However, perhaps a more important contribution of this article is the general example it sets for testing our tests in the field of cardiovascular diseases. The authors have clearly demonstrated that carrying out a prospective randomized clinical trial comparing testing modalities in a rigorous way for their impact on clinical decision making and outcomes is feasible. It requires substantial focus on an important clinical problem by a large multidisciplinary investigative team, with the patience to gather the appropriate population and observe their long-term outcomes. It also requires a belief by the investigative team in the state of clinical equipoise with regard to the question at hand; that is, that there is no clear reason to believe that one technique is clearly superior to another prior to the study.

The study by Siebelink et al. (21) is an important contribution that will, it is hoped, set a standard for carefully and critically examining other testing modalities in other clinical syndromes. At the moment and in the coming years, debates will continue to simmer regarding the appropriate use of different testing modalities for screening for early coronary disease (24), as well as other clinical scenarios such as testing after myocardial infarction. The study of Siebelink and colleagues sets an important example that should be applied to other clinical situations, so that testing our tests is seen from the template of a patient- and outcome-based perspective, rather than from a perspective of surrogate end points.

Reprint requests and correspondence: James E. Udelson, New England Medical Center, Division of Cardiology/Box 70, 750 Washington St., Boston, Massachusetts 02111. E-mail: judelson@lifespan.org.

REFERENCES

1. Fleischmann KE, Hunink MG, Kuntz KM, Douglas PS. Exercise echocardiography or exercise SPECT imaging? A meta-analysis of diagnostic test performance. JAMA 1998;280:913–20.
2. Kwok Y, Kim C, Grady D, Segal M, Redberg R. Meta-analysis of exercise testing to detect coronary artery disease in women. Am J Cardiol 1999;83:660–6.
3. The TIMI Study Group. Comparison of invasive and conservative strategies after treatment with intravenous tissue plasminogen activator in acute myocardial infarction. Results from the Thrombolysis in Myocardial Infarction (TIMI) Phase II trial. N Engl J Med 1989;320: 618–24.
4. Boden WE, O'Rourke RA, Crawford MH, et al. for the VANQWISH trial investigators. Outcomes in patients with acute non-Q-wave myocardial infarction randomly assigned to an invasive as compared with a conservative management strategy. N Engl J Med 1998;338:1785–92.
5. Dilsizian V, Bonow RO. Current diagnostic techniques of assessing myocardial viability in patients with stunned and hibernating myocardium. Circulation 1993;87:1–20.
6. Schelbert HR, Buxton D. Insights into coronary artery disease gained from metabolic imaging. Circulation 1988;78:496–505.
7. Rahimtoola SH. Hibernating myocardium has reduced blood flow at rest that increases with low-dose dobutamine. Circulation 1996;94: 3055–61.
8. Dilsizian V, Rocco TP, Freedman NM, Leon MB, Bonow RO. Enhanced detection of ischemic but viable myocardium by the reinjection of thallium after stress-redistribution imaging. N Engl J Med 1990;323:141–6.
9. Udelson JE, Coleman PS, Metherall J, et al. Predicting recovery of severe regional ventricular dysfunction: comparison of resting scintigraphy with thallium-201 and technetium-99m-sestamibi. Circulation 1994;89:2552–61.
10. Tillisch J, Brunken R, Marshall RC, et al. Reversibility of cardiac wall motion abnormalities predicted by positron tomography. N Engl J Med 1986;314:884–8.
11. Cigarroa CG, deFilippi CR, Brickner ME, Alvarez LG, Wait MA, Grayburn PA. Dobutamine echocardiography identified hibernating myocardium and predicts recovery of left ventricular function after coronary revascularization. Circulation 1993;88:430–6.
12. Gropler RJ, Bergmann SR. Myocardial viability—what is the definition? J Nucl Med 1991;32:10–12.
13. Ragosta M, Beller GA, Watson DD, Kaul S, Gimple LW. Quantitative planar rest-redistribution thallium-201 imaging in detection of myocardial viability and prediction of improvement in left ventricular function after coronary bypass surgery in patients with severely depressed left ventricular function. Circulation 1993;87:1630–41.
14. Bonow RO. Identification of viable myocardium. Circulation 1996; 94:2674–80.
15. Udelson JE. Steps forward in the assessment of myocardial viability in left ventricular dysfunction. Circulation 1998;97:833–8.
16. Samady H, Elefteriades JA, Abbott BG, Mattera JA, McPherson CA, Wackers FJ. Failure to improve left ventricular function after coronary revascularization for ischemic cardiomyopathy is not associated with worse outcome. Circulation 1999;100:1298–304.
17. Bax JJ, Poldermans D, Elhendy A, et al. Improvement of left ventricular ejection fraction, heart failure symptoms and prognosis after revascularization in patients with chronic coronary artery disease

JACC Vol. 37, No. 1, 2001
January 2001:89–92

and viable myocardium detected by dobutamine stress echocardiography. J Am Coll Cardiol 1999;34:163–9.
18. Srinivasan G, Kitsiou AN, Bacharach SL, Bartlett ML, Miller-Davis C, Dilsizian V. ^{18}F-Deoxyglucose SPECT: can it replace PET and thallium SPECT for the assessment of myocardial viability? Circulation 1998;97:843–50.
19. Bax JJ, Wijns W, Cornel JH, Visser FC, Boersma E, Fioretti PM. Accuracy of currently available techniques for prediction of functional recovery after revascularization in patients with left ventricular dysfunction due to chronic coronary artery disease: comparison of pooled data. J Am Coll Cardiol 1997;30:1451–60.
20. Allman KC, Shaw LJ, Hachamovitch R, Udelson JE. Prognostic impact of myocardial viability testing in 3088 patients: a meta-analysis (abstr). J Nucl Cardiol (in press).
21. Sibelink HJ, Blanksma PK, Crijns HJGM, et al. No difference in cardiac event-free survival between positron emission tomography-guided and single-photon emission computed tomography-guided management. A prospective, randomized comparison in patients with suspicion of jeopardized myocardium. J Am Coll Cardiol 2001;37:81–8.
22. Tamaki N, Kawamoto M, Takahashi N, et al. Prognostic value of an increase in fluorine-18 deoxyglucose uptake in patients with myocardial infarction: comparison with stress thallium imaging. J Am Coll Cardiol 1993;22:1621–7.
23. Kitsiou AN, Srinivasan G, Quyyumi AA, Summers RM, Bacharach SL, Dilsizian V. Stress-induced reversible and mild-to-moderate irreversible thallium defects: Are they equally accurate for predicting recovery of regional left ventricular function after revascularization? Circulation 1998;98:501–8.
24. O'Rourke RA, Brundage BH, Froelicher VF, et al. American College of Cardiology/American Heart Association Expert Consensus Document on electron-beam computed tomography for the diagnosis and prognosis of coronary artery disease. J Am Coll Cardiol 2000;36:326–40.